Kidney International, Vol. 57, Suppl. 74 (2000), pp. S-14–S-18

Statistical analysis, a necessity for observational studies

ROBERT A. WOLFE

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

Statistical analysis, a necessity for observational studies. Statistical models can help us to understand mechanisms that underlie associations. An example with Kt/V, gender and weight is provided.

Statistical analysis is thought of by some as a necessary evil required for publication of a paper, by others as a way to influence friends and enemies towards a particular perspective, and by still others as a confusing and irrelevant distraction from true science. Statistical tools do serve those limited purposes in the hands of some. However, in the right hands, they can prove to be of great scientific use in helping us to understand patterns in our observations, which is an important step towards understanding mechanisms underlying those data. Studies of human populations have benefited greatly from the proper use of statistical tools, in part because humans tend to be unpredictable and also because humans cannot be controlled in an experimental setting as easily as can other objects of scientific inquiry.

GOALS OF STATISTICAL ANALYSIS OF CLINICAL DATA

Perhaps the most well-known goal of statistical analysis is to account for the influence of random variation, and the tools of confidence intervals and *P*-values are useful in meeting that objective. Perhaps the most important goal, however, is to disentangle the simultaneous and combined effects of several factors on patient outcomes, in situations where those factors cannot be easily controlled experimentally.

Research questions

Clinical studies are designed to answer a variety of types of questions about the relationship of treatment methods to patient outcomes. Some examples of these questions are posed below.

Does a treatment affect patient outcomes? Unless we choose to rely upon expert opinion, the answer to this question requires that we compare patient outcomes for two treatment groups of patients. If there are differences, we need to rule out the possibility that the differences arose due to random chance. The *P*-value is one tool that is often used for this purpose. It is common practice to declare that chance is unlikely to be responsible for a difference in outcomes if the *P*-value is less than 0.05, although this choice is arbitrary. Note that this convention classifies a result as nonsignificant if the *P*-value is 0.051 and classifies a result as significant if the *P*-value is 0.049, even though a more appropriate interpretation is that the level of statistical evidence is nearly identical in these two cases. It should be recognized that this decision process is based on a social agreement to declare that if something happens infrequently (less than 1 time in 20) then it is an implausible explanation for the outcomes that we observe. More generally, a smaller *P*-value (closer to 0) indicates a more significant result, i.e., one which is harder to explain as due to chance alone. In fact, such decisions cannot be made with certainty and 5% reflects an accepted level of uncertainty in our decision.

How much does treatment affect patient outcomes? To answer this question, we must choose a criterion for evaluation of patient outcomes and quantitate the differences between treatment groups with respect to this criterion. Confidence intervals can be used to reflect the degree of uncertainty in our estimates of the size of the treatment effect. A confidence interval gives an upper and lower bound for the value of the size of the treatment effect, to reflect the uncertainty in our knowledge of its value. A 95% confidence interval will include the true value of the size of the treatment effect 95% of the time and fail to do so the remaining 5% of the time. Thus, the confidence interval is a tool with a known (but small) error rate. Unfortunately, we do not know if the result from a particular analysis yields a confidence interval that does or does not enclose the true value. The confidence interval is a tool that allows the researcher to be correct 95% of the time, but not to know when the result is right or wrong.

What are the constraints on achieving an effective treat-

Key words: dialysis dose, regression, statistical adjustment.

ment? This question can be answered by evaluating subgroups of patients. For example, separate evaluations of subgroups defined by etiology, gender, age, and level (dose) of treatment might identify the consistency, or lack of consistency, of treatment effectiveness.

What is the mechanism by which treatment affects patient outcomes? This question is very difficult to answer because "mechanism" is often based on a logical construct instead of an empiric description. However, empiric evidence can often be brought to bear by measuring the detailed responses that are implied by a particular mechanism, to determine if these stages are associated with overall differences in patient outcome. Thus, the study of patient outcomes is the most appropriate basis for evaluating a treatment, the study of intermediate measures is often more important for understanding the mechanism by which a treatment can effect patient outcomes.

Identifying causal mechanisms by exclusion

When differences are observed in patient outcomes for different treatments, we are led to another question: "To what causes might these differences be ascribed?" The answer to this question is commonly found by exclusion of alternative potential answers. Ideally, the treatments have been given to otherwise equivalent groups of patients, so that the differences in outcomes must be due, by exclusion of alternative explanations, to treatment efficacy. There are two general approaches to guaranteeing equivalent treatment groups. First, the treatment groups can be selected so that they are equivalent. Second, statistical analysis can be used to compare equivalent subgroups and to summarize the treatment comparisons across these subgroups.

Randomization of patients to treatment groups with adequate sample size assures that the treatment groups will be equivalent on average with regard to all factors, including those that are unidentified. In a well-designed randomized clinical trial, a true treatment effect and random chance are the only remaining potential explanations for differences in outcome. Thus, by ruling out random chance as an explanation, through enrollment of a large enough sample size, we are left with a true treatment effect as the only explanation for a difference in outcomes.

Other study design approaches to yield equivalent study groups include matching treatment groups with regard to important patient characteristics and use of pre- and post-treatment patient outcome measures, so that each patient serves as a self-control.

Often, however, our observations of differences in patient outcomes are based on treatment groups that differ in a variety of ways. In order to detect patient outcome differences that are due to treatment effects, rather than to treatment group characteristics, comparisons can be made within subgroups of equivalent patients. Separate analyses of patient subgroups lead to a tension between maintaining large sample sizes and making homogeneous subgroups. Regression analysis is a powerful statistical tool that can be used to aggregate treatment group differences across such subgroups.

Sources of bias

The common goal of treatment comparisons is to avoid bias. Bias can arise due to any reason, other than true treatment differences, by which patient outcomes might differ between treatment groups.

Selection bias, which leads to different types of patients in the different treatment groups, is one of the most common sources of bias. With patient selection, differences in patient outcomes could be due to either differences in treatment efficacy or to differences in the characteristics of patients in the treatment groups.

One of the most nefarious examples of selection bias is due to treatment by indication (often called "confounding by indication" or "protopathic bias"). Specific treatments are often given in response to patient symptoms. The symptoms can then affect patient outcomes, thus masking any true treatment effect. For example, patients who have been hospitalized tend to have worse outcomes than patients who are not hospitalized. However, the difference in outcomes is due to the fact that sicker patients are hospitalized while healthier patients are not. The worse outcome for hospitalized patients should not be attributed to hospitalization as a treatment.

Another common source of bias can arise from different standards of reporting for patient outcomes in different treatment groups. For example, ascertainment bias can occur because rates of disease are higher for patients receiving medical treatment than for the general population, so higher reported cancer rates might be expected among end-stage renal disease patients than in the general population due to the fact that they are under medical supervision.

Treatment groups that differ with regard to time of treatment are especially prone to bias. If the criteria for patient entry into treatment have changed over time, then this lead-time bias will cause an apparent change in patient outcomes as a result. At least part of the reduction in mortality among cancer patients in recent years is due to the fact that cancers are being treated at less advanced stages than they were in previous years due to earlier detection of cancers.

Treatment group effects on patient outcomes might be due to unexpected mechanisms. Planned changes in a treatment protocol might also be accompanied by other changes that can affect patient outcomes. New treatments that are pioneered by exceptional facilities might have exceptional outcomes because of other characteristics at the facilities and might not be due to the new

Table 1. Males have lower Kt/V and higher weight than females

	Kt/V	Weight (kg)
Male	1.14 ± 0.25	73.2 ± 14
Female	1.27 ± 0.30	63.7 ± 17
Delta	0.127	9.5
T	5.96	−7.81
P-value	<0.0001	<0.0001

treatment. It is often difficult to generalize results from a well-controlled protocol to a wider diversified setting in which protocol deviations (such as patient noncompliance) are more common.

Regression analysis example

Clinical studies are often used to compare patient outcomes for different treatment groups. With observational studies, regression analysis offers a method to answer the question: How big would the treatment difference have been, had the treatment groups been equivalent with respect to other factors? This section presents a detailed interpretation of a regression analysis to show how such an analysis can estimate the effects that two different factors have on an outcome. This example shows how the dose of dialysis, as measured by delivered Kt/V, differs by patient gender and weight. These data are from a United States Renal Data System (USRDS) study of a random sample of patients in the United States in 1993 [1]. The example is linked to clinical studies of patient outcomes by comparing two groups (genders) with respect to an outcome (Kt/V) while accounting for differences in patient characteristics (weight).

The single pool dose of dialysis was evaluated based on the pre-/post-dialysis BUN for a sample of 680 patients, along with patient gender and predialysis weight. Table 1 summarizes the differences in weight and Kt/V for males and females. The average Kt/V is 0.127 units higher for females than for males ($P < 0.0001$). This difference is not due to chance, since the *P*-value is so significant, but it is unclear which aspects of gender might "cause" the difference in average Kt/V. One possible mechanism could be that males are heavier than females, on average, and heavier people tend to receive lower levels of Kt/V. Table 1 shows that in fact, the males are 9.5 kg heavier than the females on average ($P < 0.0001$). In order to be responsible for the difference in Kt/V, weight must also be associated with Kt/V.

Figure 1 plots the relationship between weight and Kt/V, along with a best fit (least squares) regression line that approximates the average Kt/V in the weight range shown. The data points indicate that average Kt/V has an approximately linear relationship to weight. The regression coefficient (β) indicates that for every kilogram heavier, patients receive 0.0074 units less of Kt/V, on average. The *P*-value ($P < 0.001$) for testing if the regression coefficient is equal to 0 indicates that chance is an unlikely explanation for the strong association seen in these data. The R-square statistic that indicates that 18.6% of the variability among patients in Kt/V can be attributed to differences in their weights. Could weight difference between males and females be solely responsible for the lower Kt/V observed among males?

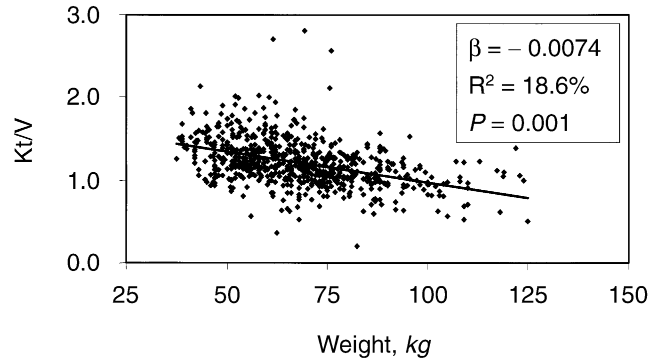

Fig. 1. Kt/V is negatively associated with weight.

Table 2. Average Kt/V is lower for males in most weight ranges

	Female		Male	
Weight range (kg)	Kt/V	*N*	Kt/V	*N*
<50	1.38	75	1.45	4
50–59	1.35	103	1.27	49
60–69	1.27	72	1.23	92
70–79	1.18	45	1.13	88
80–89	1.11	37	1.04	47
90–99	1.13	15	0.98	21
100–109	0.96	5	0.86	15
110+	0.97	7	0.97	5

Table 2 shows that males have lower Kt/V in every weight subgroup shown except for the two most extreme groups, where random variation in small sample sizes may obscure the underlying pattern. In most of the weight ranges, females have higher average Kt/V, by between 0.04 and 0.15 units. Several large audiences have studied this table and concluded that for males and females in the same weight range, the average Kt/V among females is about 0.06 units higher than for males, overall. Random chance undoubtedly contributes to the variation in the Kt/V difference among weight groups reported in this table for males and females.

Further, examining each gender separately shows that the average Kt/V tends to decrease with higher weight. Ignoring the most extreme groups, the middle six groups span a 50-kg range with a corresponding change in average Kt/V very close to 0.4 units. This could be summarized as a decrease of about 0.008 units of Kt/V for each kg heavier, which is similar to the results of the regression analysis above.

Table 3. Regress Kt/V on weight and sex

Variable	β	*P*-value
Intercept	1.71	0.0001
Weight	−0.0069	0.0001
Male	−0.062	0.003

$R^2 = 19.6\%$
$N = 680$

Table 4. Examine the independent contributions of weight and sex to Kt/V

	Average Kt/V
Female, 70 kg	$1.23 = 1.71 - 0.062 \times 0 - 0.0069 \times 70$
Male, 70 kg	$1.17 = 1.71 - 0.062 \times 1 - 0.0069 \times 70$
Δ Adjust.	$0.062 = 0 + 0.062 + 0$

Table 5. Examine the independent contributions of weight and sex to Kt/V

	Average Kt/V
Female, 63.7 kg	$1.27 = 1.71 - 0.062 \times 0 - 0.0069 \times 63.7$
Male, 73.2 kg	$1.14 = 1.71 - 0.062 \times 1 - 0.0069 \times 73.2$
Δ crude	$0.127 = 0 + 0.062 + 0.0069 \times 9.5$

A regression model of these same data yields the results in Table 3. The regression coefficient indicates that within gender group the average Kt/V is lower by −0.0069 for each kg heavier, and that for males and females of the same weight the average Kt/V is 0.062 units lower for males than for females. These results are in good agreement with the interpretation of Table 2.

The results of Table 3 yield an equation for the average Kt/V for any given gender and weight combination: Kt/V = 1.71 − 0.0069 × weight (kg) for females and Kt/V = 1.71 − 0.062 − 0.0069 × weight (kg) for males. Two calculations are given below to show the interpretation of these equations.

Table 4 shows the calculation of average Kt/V for 70 kg males and females. For 70 kg males and females, respectively, these two equations yield average Kt/V estimates of 1.23 and 1.17, which differ by 0.062. Note that the same difference of 0.062 units would result for males and females of any weight, so long as they were the same weight. This is the difference in average Kt/V for males and females, adjusted for weight.

Figure 2 shows the calculation of average Kt/V for males and females of average weight for their gender. The equation for females yields an average Kt/V of 1.27 for females of weight 63.7 kg. Similarly, the equation for males yields an average Kt/V of 1.14 for males of weight 73.2 kg. Both equations replicate the observed average Kt/V for each gender at the average weight for each gender (Table 1). The bottom line of the figure shows

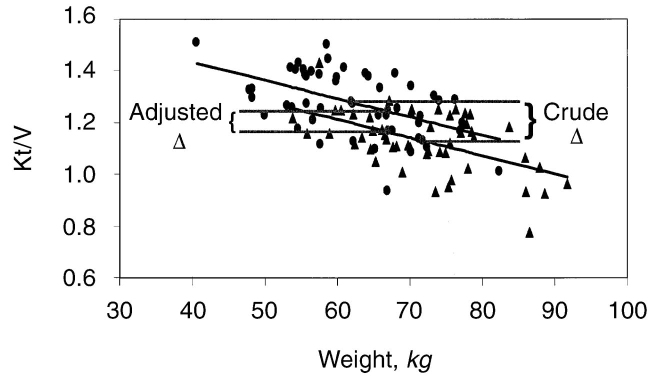

Fig. 2. Kt/V by weight and sex. ▲, men; ●, women.

the difference in the calculated weights for these two groups and shows that the total difference of 0.127 units between males and females can be split into two components: 0.062 units due to gender alone, and 0.065 units due to the difference of 9.5 kg in the average weight of these two groups.

DISCUSSION

In the example above, the difference in outcome (Kt/V) between two groups (by gender) was exaggerated by another factor (weight). Another nonrandomized example will be given here to show that differences between two groups can also be obscured by another factor. Many recent studies have compared differences in mortality between high and low Kt/V treatment groups. Due to the dialysis dosing practices in the United States, larger patients tend to receive lower Kt/V than smaller patients, on average. Thus, mortality comparisons, without adjustment for patient size, of high and low dialysis dose groups show the net result of both different doses and of different sizes of patients. A recent study [2] has shown that larger patients tend to have a survival advantage over smaller patients. In this situation, any benefit of higher dose of dialysis on patients in a high dose group would tend to be canceled partly by the adverse effect of their smaller body size. Either statistical adjustments (regression analysis), or subgroup analysis can be used to separate the unique contributions of dose of dialysis and patient size on patient mortality.

Only the simplest applications of regression analysis have been discussed here. As with many sophisticated tools, regression analysis can be easily misused, even in the hands of a well-meaning analyst. Detailed review of tables and graphs can help to identify many of the pitfalls of regression analysis. There are many instructive texts that show how to use this tool effectively [3–5].

CONCLUSION

The results of observational studies can be very useful in the identification of treatment effects on patient outcomes. The difficult question after identifying a difference between treatment groups is in understanding the cause of the difference. Is it due to random fluctuation, to bias, or to the treatments themselves? Accounting for random fluctuations is an important step in nearly any analysis of data from a clinical study. Both *P*-values and confidence intervals are useful for this purpose. Randomization is an excellent design approach for avoiding bias. Regression analysis can help to quantitate the effects that each of several factors has had on patient outcomes. Regression analysis can help to answer the question: How big would the difference in outcomes have been, had the treatment groups been otherwise equivalent?

Reprint requests to Robert A. Wolfe, Ph.D, Kidney Epidemiology and Cost Center, 315 W. Huron, Suite 240, Ann Arbor, Michigan 48103, USA.
E-mail: bobwolfe@umich.edu

REFERENCES

1. Held PJ, Port FK, Webb RL, Wolfe RA, Bloembergen WE, Turenne MN, Holzman E, Ojo AO, Young EW, Mauger EA, Tedeschi PJ, Stannard DC, Strawderman RL, Carroll CE, Levine GN, Wood CL, Smith DA, Jones CA, Greer JW, Hill DJ, Ketz L-AD, Agodoa LYC: United States Renal Data System 1996 Annual Data Report. An Introduction. *Am J Kidney Dis* 26 (Suppl. 2):S1–S3 and 1–186, 1996
2. Wolfe RA, Ashby VB, Daugirdas JT, Agodoa LYC, Jones CA, Port FK: Body size, dose of hemodialysis and mortality. *Am J Kidney Dis*, in press
3. Allen DM, Cady FB: *Analyzing Experimental Data by Regression.* Belmont, CA, Lifetime Learning Publications, 1982
4. Daniel C, Wood F: *Fitting Equations to Data.* New York, John Wiley and Sons, 1980
5. Moore DS, McCabe GP: *Introduction to the Practice of Statistics* (3rd edn). New York, Freeman, 1999